TOEFL® Reference List

Abraham, R. G., & Plakans, B. S. (1988). Evaluating a screening/training program for NNS teaching assistants. *TESOL Quarterly, 22*, 505–508. https://doi.org/10.2307/3587294

Aertselaer, J. N.-V. (2013). Contextualizing EFL argumentation writing practices within the Common European Framework descriptors. *Journal of Second Language Writing, 22*, 198–209. https://doi.org/10.1016/j.jslw.2013.03.010

Al-Hawamdeh, R. F., & Al-Khanji, R. (2017). The effects of motivation and other factors on second language acquisition: A case study on achieving advanced oral proficiency in English. *Arab World English Journal, 8*(1), 165–178. https://doi.org/10.24093/awej/vol8no1.12

Alavi, S. M., & Bordbar, S. (2012a). Are they right participants for the right strategies? A case study in the role of levels of language ability in strategy use in reading section of TOEFL iBT®. Theory & Practice in Language Studies, 2, 877–886. https://doi.org/10.4304/tpls.2.5.877-886

Alavi, S. M., & Bordbar, S. (2012b). A closer look at reading strategy use in reading section of TOEFL iBT. Theory & Practice in Language Studies, 2, 450–460. https://doi.org/10.4304/tpls.2.3.450-460

Alderman, D. L. (1981). *Language proficiency as a moderator variable in testing academic aptitude* (Research Report No. RR-81-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01268.x

Alderman, D. L., & Holland, P. W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language™* (Research Report No. RR-81-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01251.x

Alderson, J. C. (2009). Test of English as a Foreign Language™: Internet-based Test (TOEFL® iBT). *Language Testing, 26*, 621–631. https://doi.org/10.1177/0265532209346371

Almond, R. G., & Mislevy, R. J. (1998). *Graphical models and computerized adaptive testing* (Research Report No. RR-98-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01753.x

Altinsoy, E., & Boyraz, S. (2017). The relationship between metalinguistic knowledge and speaking anxiety of ELT students. *International Journal of Language Academy, 5*, 151–158. https://doi.org/10.18033/ijla.3653

Aminloo, M. S. (2013). The effect of collaborative writing on EFL learners writing ability at elementary level. *Journal of Language Teaching & Research, 4*, 801–806. https://doi.org/10.4304/jltr.4.4.801-806

Amiryousefi, M., & Tavakoli, M. (2011). The relationship between test anxiety, motivation and MI and the TOEFL iBT reading, listening and writing scores. *Procedia—Social and Behavioral Sciences, 15*, 210–214. https://doi.org/10.1016/j.sbspro.2011.03.075

Anderson-Hsieh, J. (1990). Teaching suprasegmentals to international teaching assistants using field-specific materials. *English for Specific Purposes, 9*, 195–214. https://doi.org/10.1016/0889-4906(90)90013-3

Anderson-Hsieh, J. (1992). Using electronic visual feedback to teach suprasegmentals. *System, 20*, 51–62. https://doi.org/10.1016/0346-251X(92)90007-P

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*, 561–613. https://doi.org/10.1111/j.1467-1770.1988.tb00167.x

Angelis, P. J., Swinton, S. S., & Cowell, W. R. (1979). *The performance of non-native speakers of English on TOEFL® and verbal aptitude tests* (Research Report No. RR-79-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1979.tb01175.x

Angoff, W. H. (1989). *Context bias in the Test of English as a Foreign Language* (Research Report No. RR-89-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1989.tb00336.x

Arshavskaya, E. (2015). International Teaching Assistants' Experiences in the U.S. Classrooms: Implications for Practice. *Journal of the Scholarship of Teaching and Learning, 15*(2), 56–69. https://doi.org/10.14434/josotl.v15i2.12947

Arshavskaya, E. (2016). Using reflective dialogic blogs with international teaching assistants: Rationale, context, and findings. *Writing & Pedagogy, 8*, 333–359. https://doi.org/10.1558/wap.26725

Asmani, A. B. (2014). Correlative analysis of TOEFL iBT scores of listening skill versus scores of Business English speaking skill among Binus University sophomores. *Lingua Cultura, 8*, 85–94. https://doi.org/10.21512/lc.v8i2.447

Attali, Y. (2007). *Construct validity of e-rater® in scoring TOEFL® essays* (Research Report No. RR-07-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y. (2011). *Automated subscores for TOEFL iBT® independent essays* (Research Report No. RR-11-39). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02275.x

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment, 10*(3), 1–6.

Attali, Y., Powers, D. E., Tannenbaum, R. J., Wylie, E. C., Kim, S., Walker, M. E., . . . Briller, V. (2009). ETS Research Spotlight D. Eignor (Ed.) ETS Research Spotlight. Retrieved from http://www.ets.org/Media/Research/pdf/SPOTLIGHT2.pdf

Attali, Y., & Sinharay, S. (2015). *Automated trait scores for TOEFL® writing tasks* (Research Report No. RR-15-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12061

Babenko, T. (2013). Speaking and writing strategies for TOEFL® iBT. *TESOL Journal, 4*(1), 194–196. https://doi.org/10.1002/tesj.70

Bailey, K. M. (1999). *Washback in language testing* (Research Memorandum No. RM-99-04). Princeton, NJ: Educational Testing Service.

Banerjee, J., & Clapham, C. (2003). Test review: The TOEFL CBT (Computer-based test). *Language Testing, 20*, 111–123. https://doi.org/10.1191/0265532203lt246xx

Banzina, E., Hewitt, L. E., & Dilley, L. C. (2014). Using synchronous speech to facilitate acquisition of English rhythm: A small scale study. *EuroAmerican Journal of*

Applied Linguistics and Languages, 1(1), 69–84. https://doi.org/10.21283/2376905X.1.9

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL iBT writing tasks. *Language Testing, 31*, 241–259. https://doi.org/10.1177/0265532213509810

Barkaoui, K. (2015). *Test takers' writing activities during the TOEFL iBT® writing tasks: A stimulated recall study* (Research Report No. RR-15-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12050

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *Modern Language Journal, 100*, 320–340. https://doi.org/10.1111/modl.12316

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics, 34*, 304–324. https://doi.org/10.1093/applin/ams046

Barnes, M. (2016). The washback of the TOEFL iBT in Vietnam. *Australian Journal of Teacher Education, 41*(7), 158–174. https://doi.org/10.14221/ajte.2016v41n7.10

Barnes, M. (2017). Washback: Exploring what constitutes "good" teaching practices. *Journal of English for Academic Purposes, 30*, 1–12. https://doi.org/10.1016/j.jeap.2017.10.003

Baron, P. A., & Papageorgiou, S. (2014). *Mapping the TOEFL® Primary™ Test onto the Common European Framework of Reference* (Research Memorandum No. RM-14-05). Princeton, NJ: Educational Testing Service.

Baron, P. A., & Papageorgiou, S. (2016). *Setting language proficiency score requirements for English-as-a-second-language placement decisions in secondary education* (Research Report No. RR-16-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12102

Baron, P. A., & Tannenbaum, R. J. (2011). *Mapping the TOEFL Junior® test onto the Common European Framework of Reference* (Research Memorandum No. RM-11-07). Princeton, NJ: Educational Testing Service.

Beardsmore, H. B., & Renkin, A. (1971). Test of spoken English™. *IRAL: International Review of Applied Linguistics in Language Teaching, 9*, 1–11. https://doi.org/10.1515/iral.1971.9.1.1

Bejar, I. I. (1985a). *A preliminary study of raters for the Test of Spoken English™* (Research Report No. RR-85-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00090.x

Bejar, I. I. (1985b). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Research Report No. RR-85-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00096.x

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL® 2000 listening framework: A working paper* (Research Memorandum No. RM-00-07). Princeton, NJ: Educational Testing Service.

Ben-David, M. F., Klass, D. J., Boulet, J., Champlain, A. D., King, A. M., Pohl, H. S., & Gary, N. C. (1999). The performance of foreign medical graduates on the National Board of Medical Examiners (NBME) standardized patient examination

prototype: A collaborative study of the NBME and the educational Commission for Foreign Medical Graduates (ECFMG). *Medical Education, 33*, 439–446. https://doi.org/10.1046/j.1365-2923.1999.00368.x

Bhat, S., & Yoon, S.-Y. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication, 67*, 42–57. https://doi.org/10.1016/j.specom.2014.09.005

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. In C. Meyer & P. Leistyna (Eds.), *Corpus analysis: Language structure and language use* (pp. 47–70). Amsterdam, The Netherlands: Rodopi.

Biber, D. (2006a). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5*, 97–116. https://doi.org/10.1016/j.jeap.2006.05.001

Biber, D. (2006b). *University language: A corpus-based study of spoken and written registers*. Amsterdam, The Netherlands: John Benjamins.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly, 36*, 9–48. https://doi.org/10.2307/3588359

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., . . . Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL® 2000 spoken and written academic language corpus* (Research Memorandum No. RM-04-03). Princeton, NJ: Educational Testing Service.

Biber, D., Csomay, E., Jones, J. K., & Keck, C. (2004a). A corpus linguistic investigation of vocabulary-based discourse units in university registers. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multi-dimensional perspective* (pp. 53–72). Amsterdam, The Netherlands: Rodopi.

Biber, D., Csomay, E., Jones, J. K., & Keck, C. (2004b). Vocabulary-based discourse units in university registers. In A. Partington, J. Morley, & L. Haarman (Eds.), *Corpora and discourse* (pp. 23–40). Bern, Switzerland: Peter Lang.

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT® test: A lexico-grammatical analysis* (Research Report No. RR-13-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam: Task types and proficiency levels. *Applied Linguistics, 37*, 639–668. https://doi.org/10.1093/applin/amu059

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis* (Research Report No. RR-11-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Biber, D., Reppen, R., & Staples, S. (2017). Exploring the relationship between TOEFL iBT scores and disciplinary writing performance. *TESOL Quarterly, 51*, 948–960. https://doi.org/10.1002/tesq.359

Boldt, R. F. (1988). *Latent structure analysis of the Test of English as a Foreign Language* (Research Report No. RR-88-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00283.x

Boldt, R. F. (1991). *Cross-validation of a proportional item response curve model* (Research Report No. RR-91-33). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01400.x

Boldt, R. F. (1992). *Reliability of the Test of Spoken English revisited* (Research Report No. RR-92-52). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01483.x

Boldt, R. F. (1994). *Simulated equating using several item response curves* (Research Report No. RR-93-57). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01568.x

Boldt, R. F., & Courtney, R. G. (1997). *Survey of standards for foreign student applicants* (Research Report No. RR-97-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1997.tb01728.x

Boldt, R. F., & Freedle, R. O. (1996). *Using a neural net to predict item difficulty* (Research Report No. RR-96-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1996.tb01709.x

Boldt, R. F., Larsen-Freeman, D., Reed, M. S., & Courtney, R. G. (1992). *Distributions of ACTFL ratings by TOEFL® score ranges* (Research Report No. RR-92-59). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01490.x

Boldt, R. F., & Oltman, P. K. (1993). *Multimethod construct validation of the Test of Spoken English* (Research Report No. RR-93-58). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01569.x

Boraie, D., Arrigoni, E., & Moos, J. (2017). A survey of English language proficiency requirements for admission to English-medium universities in Arabic-speaking countries. In A. Gebril (Ed.), *Applied linguistics in the Middle East and North Africa* (pp. 228–247). Amsterdam, the Netherlands: John Benjamins.

Boyd, F. A. (1989). Developing presentation skills: A perspective derived from professional education. *English for Specific Purposes, 8*, 195–203. https://doi.org/10.1016/0889-4906(89)90030-6

Breland, H., & Lee, Y.-W. (2007). Investigating uniform and non-uniform gender DIF in computer-based ESL writing assessment. *Applied Measurement in Education, 20*, 377–403. https://doi.org/10.1080/08957340701429652

Breland, H., Lee, Y.-W., & Muraki, E. (2004a). *Comparability of TOEFL® CBT writing prompts: Response mode analyses* (Research Report No. RR-04-23). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01950.x

Breland, H., Lee, Y.-W., & Muraki, E. (2004b). Comparability of TOEFL CBT essay prompts: Responsenmode analyses. *Educational and Psychological Measurement, 65*, 577–595.

Breland, H., Lee, Y.-W., Najaran, M., & Muraki, E. (2004). *An analysis of TOEFL® CBT writing prompt difficulty and comparability for different gender groups* (Research Report No. RR-04-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01932.x

Breland, H. M., Kubota, M. Y., Nickerson, K., Trapani, C. S., & Walker, M. E. (2004). *New SAT Writing Prompt Study: Analyses of Group Impact and Reliability* (Research Report No. RR-04-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01930.x

Bresnahan, M. I., & Kim, M. S. (1993a). The impact of positive and negative messages on change in attitude toward international teaching assistants. *Folia Linguistica, 27*, 347–363. https://doi.org/10.1515/flin.1993.27.3-4.347

Bresnahan, M. I., & Kim, M. S. (1993b). Predictors of recetivity and resistance toward international teaching assistants. *Journal of Asian Pacific Communication (Multilingual Matters), 4*, 3–14.

Bresnahan, M. J., & Cai, D. H. (2000). From the Other Side of the Desk: Conversations with International Students about Teaching in the U.S. *Qualitative Research Reports in Communication, 1*(4), 65–75.

Bridgeman, B. (2016). Can a two-question test be reliable and valid for predicting academic outcomes? *Educational Measurement: Issues & Practice, 35*(4), 21–24. https://doi.org/10.1111/emip.12130

Bridgeman, B., & Carlson, S. (1983). *Survey of academic writing tasks required of graduate and undergraduate foreign students* (Research Report No. RR-83-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1983.tb00018.x

Bridgeman, B., Cho, Y., & DiPietro, S. (2016). Predicting grades from an English language assessment: The importance of peeling the onion. *Language Testing, 33*, 307–318. https://doi.org/10.1177/0265532215583066

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT® speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*, 91–108. https://doi.org/10.1177/0265532211411078

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*, 27–40. https://doi.org/10.1080/08957347.2012.635502

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly, 11*, 353–373. https://doi.org/10.1080/15434303.2014.947532

Brooks, L., & Swain, M. (2015). Students' voices: The challenge of measuring speaking for academic contexts. In B. Spolsky, O. Inbar, & M. Tannenbaum (Eds.), *Challenges for language education and policy: Making space for people* (pp. 65–80). New York, NY: Routledge.

Brown, G. (1981). Teaching the Spoken Language. *Studia Linguistica, 35*, 166–182. https://doi.org/10.1111/j.1467-9582.1981.tb00708.x

Brutten, S. R., Angelis, P. J., & Perkins, K. (1985). Music and memory: predictors for attained ESL oral proficiency. *Language Learning, 35*, 299–313. https://doi.org/10.1111/j.1467-1770.1985.tb01030.x

Burstein, J. (2012). Fostering best practices in writing assessment and Instruction with R-rater. In N. Elliot & L. C. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 203–217). Creskill, NJ: Hampton Press.

Burstein, J. C., Kaplan, R. M., Rohen-Wolff, S., Zuckerman, D. I., & Chi, L. (1999). *A review of computer-based speech technology for TOEFL® 2000* (Research Memorandum No. RM-99-05). Princeton, NJ: Educational Testing Service.

Butler, F., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL® 2000 speaking framework: A working paper* (Research Memorandum No. RM-00-06). Princeton, NJ: Educational Testing Service.

Caldwell, K., & Samuel, C. (2001). Test of Spoken English (*TSE®*). *Canadian Modern Language Review, 58*, 319–326.

Carey, P. (1996). *A review of psychometric and consequential issues related to performance assessment* (Research Memorandum No. RM-96-03). Princeton, NJ: Educational Testing Service.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English* (Research Report No. RR-85-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00106.x

Carrell, P. L. (2007). *Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks* (Research Report No. RR-07-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02043.x

Carrell, P. L., Dunkel, P. A., & Mollaun, P. (2002). *The effects of notetaking, lecture length and topic on the listening component of TOEFL® 2000* (Research Memorandum No. RM-02-04). Princeton, NJ: Educational Testing Service.

Carrell, P. L., Dunkel, P. A., & Mollaun, P. (2004). The effects of notetaking, lecture length, and topic on a computer-based test of ESL listening comprehension. *Applied Language Learning, 14*, 83–105.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*, 383–391. https://doi.org/10.1111/j.0083-2919.2005.00419.x

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(3), 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chapelle, C., Grabe, W., & Berns, M. (1997). *Communicative language proficiency: Definition and implications for TOEFL® 2000* (Research Memorandum No. RM-97-03). Princeton, NJ: Educational Testing Service.

Chen, J., & Sheehan, K. M. (2015). *Analyzing and comparing reading stimulus materials across the TOEFL® Family of Assessments* (Research Report No. RR-15-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12055

Chiang, S.-Y. (2009a). Dealing with Communication Problems in the Instructional Interactions between International Teaching Assistants and American College Students. *Language and Education, 23*, 461–478. https://doi.org/10.1080/09500780902822959

Chiang, S.-Y. (2009b). Mutual understanding as a procedural achievement in intercultural interaction. *Intercultural Pragmatics, 6*, 367–394. https://doi.org/10.1515/IPRG.2009.019

Chiang, S.-Y. (2011). Pursuing a response in office hour interactions between US college students and international teaching assistants. *Journal of Pragmatics, 43*, 3316–3330. https://doi.org/10.1016/j.pragma.2011.07.001

Chiang, S.-Y. (2016). 'Is This What You're Talking About?': Identity Negotiation in International Teaching Assistants' Instructional Interactions with U.S. College Students. *Journal of Language, Identity & Education, 15*, 114–118. https://doi.org/10.1080/15348458.2016.1137726

Chiang, S.-Y., & Mi, H.-F. (2008). Reformulation as a strategy for managing 'understanding uncertainty' in office hour interactions between international teaching assistants and American college students. *Intercultural Education, 19*, 269–281. https://doi.org/10.1080/14675980802078640

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing, 29*, 421–442. https://doi.org/10.1177/0265532211430368

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2016). *Designing the TOEFL Primary® Tests* (Research Memorandum No. RM-16-02). Princeton, NJ: Educational Testing Service.

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2017). Designing the TOEFL Primary Tests. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 41–58). New York, NY: Routledge.

Cho, Y., Rijmen, F., & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT™ integrated writing tasks. *Language Testing, 30*, 513–534. https://doi.org/10.1177/0265532213478796

Cho, Y., & So, Y. (2014). *Construct-irrelevant factors influencing young English as a Foreign Language (EFL) learners' perceptions of test task difficulty* (Research Memorandum No. RM-14-04). Princeton, NJ: Educational Testing Service.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater®'s performance on TOEFL® essays* (Research Report No. RR-04-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01931.x

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing, 27*, 419–436. https://doi.org/10.1177/0265532210364391

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*, 295–320. https://doi.org/10.1191/0265532203lt258oa

Choi, I. (2017). Empirical profiles of academic oral English proficiency from an international teaching assistant screening test. *Language Testing, 34*, 49–82. https://doi.org/10.1177/0265532215601881

Choi, I., & Papageorgiou, S. (2014). *Monitoring students' progress in English language skills using the TOEFL ITP® assessment series* (Research Memorandum No. RM-14-11). Princeton, NJ: Educational Testing Service.

Chumpavan, S., Lorber, M., Al-Bataineh, A., & Abu Al-Rub, M. (2008). Global connections: Using technology to teach a second language. *International Journal of Learning, 15*, 133–147. https://doi.org/10.18848/1447-9494/CGP/v15i02/45633

Chun, C. W. (2006). COMMENTARY: An Analysis of a Language Test for Employment: The Authenticity of the PhonePass Test. *Language Assessment Quarterly, 3*, 295–306. https://doi.org/10.1207/s15434311laq0303_4

Chyn, S., Tang, K. L., & Way, W. D. (1994). *An investigation of IRT-based assembly of the TOEFL® test* (Research Report No. RR-94-38). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01611.x

Clara, H. (2011). Phonological analysis of university students' spoken discourse. *Humaniora, 2*(1), 77–82. https://doi.org/10.21512/humaniora.v2i1.2951

Clark, J. L. D. (1977). *The performance of native speakers on English on the Test of English as a Foreign Language™* (TOEFL Research Report No. 01). Princeton, NJ: Educational Testing Service.

Clark, J. L. D., & Swinton, S. S. (1979). *An exploration of speaking proficiency measures in the TOEFL context* (Research Report No. RR-79-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1979.tb01176.x

Clark, J. L. D., & Swinton, S. S. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (Research Report No. RR-80-33). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1980.tb01230.x

Cohen, A. D., & Upton, T. A. (2004). Strategies in responding to the next generation TOEFL reading tasks. *Language Testing Update, 35*, 53–55.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (Research Report No. RR-06-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02012.x

Cohen, A. D., & Upton, T. A. (2007). I want to go back to the text: Response strategies on the reading subtest of the New TOEFL®. *Language Testing, 24*, 209–250. https://doi.org/10.1177/0265532207076364

Compton, L. K. L. (2007). The Impact of Content and Context on International Teaching Assistants' Willingness to Communicate in the Language Classroom. *TESL-EJ, 10*(4), 1–20.

Constantinides, J. C. (1989). ITA Training Programs. *New Directions for Teaching and Learning*, *39*, 71–77. https://doi.org/10.1002/tl.37219893908

Corrigan, P. C. (2015). English For the Medium of Instruction (EFMI) at a University in Hong Kong. *IAFOR Journal of Education, 3*(2), 158–170. https://doi.org/10.22492/ije.3.2.10

Coşkun, A., & Ghaemi, H. (2015). Integrating Technologically-Enhanced Self-Regulated Strategies into Writing English as a Foreign Language Classes. *International Online Journal of Educational Sciences, 7*(2), 1–14. https://doi.org/10.15345/iojes.2015.02.006

Crossley, S., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly, 11*, 250–270. https://doi.org/10.1080/15434303.2014.926905

Crossley, S. A., Kyle, K., Varner, L., Gou, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment, 7*(1). Retrieved from http://journalofwritingassessment.org/

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology, 17*, 171–192.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *Modern Language Journal, 99*(1), 80–95. https://doi.org/10.1111/modl.12185

Crumley, H. (2010). Instructional Technology in International Teaching Assistant (ITA) Programs. *CALICO Journal*, 409-431. https://doi.org/10.11139/cj.27.2.409-431

Cumming, A. (2010). Review of Building a validity argument for the test of English as a foreign language™. *Language Testing, 27*, 286–288. https://doi.org/10.1177/0265532210270020902

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*, 159–197. https://doi.org/10.1191/0265532204lt278oa

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL®* (Research Memorandum No. RM-04-05). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®* (Research Report No. RR-05-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01990.x

Cumming, A., Kantor, R., Baba, K., Erdosy, M. U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for the next generation TOEFL. *Assessing Writing, 10*, 5–43. https://doi.org/10.1016/j.asw.2005.02.001

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL® essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (Research Memorandum No. RM-01-04). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*, 67–96. https://doi.org/10.1111/1540-4781.00137

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL® 2000 writing framework: A working paper* (Research Memorandum No. RM-00-05). Princeton, NJ: Educational Testing Service.
Davies, C. E., Tyler, A., & Koran, J. J. J. (1989). Face-to-face with english speakers: An advanced training class for international teaching assistants. *English for Specific Purposes, 8*, 139–153. https://doi.org/10.1016/0889-4906(89)90026-4
Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*, 117–135. https://doi.org/10.1177/0265532215582282
Davis, W. E. (1991). Comparing language proficiency scores and student evaluations to determine policy for international teaching assistants. *College Student Journal, 25*, 489–495.
Deane, P., & Gurevich, O. (2008). *Applying content similarity metrics to corpus data: Differences between native and non-native speaker responses to a TOEFL® integrated writing prompt* (Research Report No. RR-08-51). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02137.x
DeLuca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: An exploratory study on the TOEFL iBT. *System, 41*, 663–676. https://doi.org/10.1016/j.system.2013.07.010
DeMauro, G. E. (1992). *An investigation of the appropriateness of the TOEFL® test as a matching variable to equate TWE® topics* (Research Report No. RR-92-26). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01457.x
DeMauro, G. (1992). Examination of the relationships among TSE, TWE and TOEFL scores. *Language Testing, 9*, 149–161. https://doi.org/10.1177/026553229200900203
Deroey, K. L. B. (2015). Marking Importance in Lectures: Interactive and Textual Orientation. *Applied Linguistics, 36*, 51–72. https://doi.org/10.1093/applin/amt029
Deroey, K. L. B., & Taverniers, M. (2012). Just remember this: Lexicogrammatical relevance markers in lectures. *English for Specific Purposes, 31*, 221–233. https://doi.org/10.1016/j.esp.2012.05.001
Dick, R. C., & Robinson, B. M. (1994). Oral English proficiency requirements for ITAs in U. S. colleges and universities: An issue in speech communication. *Journal of the Association for Communication Administration, 2*, 77–86.
Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (Research Memorandum No. RM-97-01). Princeton, NJ: Educational Testing Service.
Douglas, D., & Myers, C. (1989). TAs on TV: Demonstrating communication strategies for international teaching assistants. *English for Specific Purposes, 8*, 169–179. https://doi.org/10.1016/0889-4906(89)90028-8
Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (Research Memorandum No. RM-97-02). Princeton, NJ: Educational Testing Service.
Du Steinberg, W. (2007). The ITA Program: An Academic Bridging Program for the Changing Demographics on North American Campuses. *Journal of Continuing*

Higher Education, 55(3), 31–37. https://doi.org/10.1080/07377366.2007.10400128
Duran, R. P., Canale, M., Penfield, J., Stansfield, C. W., & Liskin-Gasparro, J. E. (1985). *TOEFL® from a communicative viewpoint on language proficiency: A working paper* (Research Report No. RR-85-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00093.x
Eignor, D., Taylor, C., Kirsch, I., & Jamieson, J. (1998). *Development of a scale for assessing the level of computer familiarity of TOEFL® examinees* (Research Report No. RR-98-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01756.x
Enright, M. K. (2004). Research issues in high-stakes communicative language testing: Reflections on TOEFL's new directions. *TESOL Quarterly, 38*, 147–151. https://doi.org/10.2307/3588266
Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL® 2000 reading framework: A working paper* (Research Memorandum No. RM-00-04). Princeton, NJ: Educational Testing Service.
Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater scoring. *Language Testing, 27*, 317–334. https://doi.org/10.1177/0265532210363144
Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (Research Report No. RR-03-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01909.x
Erfani, S. S. (2012). A comparative washback study of IELTS and TOEFL iBT on teaching and learning activities in preparation courses in the Iranian context. *English Language Teaching, 5*(8), 185–195.
ETS. (2010a). *Linking TOEFL iBT™ Scores to IELTS Scores - A Research Report*, 1-17. Retrieved from http://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf
ETS. (2010b). TOEFL iBT™ Test Framework and Test Development. *TOEFL iBT™ Research Insight, 1*(1), 1-8. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_insight.pdf
ETS. (2010c). TOEFL Research. *TOEFL iBT™ Research Insight*, *1*(2), 1-8. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v2.pdf
ETS. (2011a). Reliability and Comparability of TOEFL iBT™ Scores. *TOEFL iBT™ Research Insight, 1*(3), 1-8. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf
ETS. (2011b). Validity Evidence Supporting the Interpretation and Use of TOEFL iBT™ Scores. *TOEFL iBT™ Research Insight*, *1*(4), 1-12. Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf
Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). *Automated scoring for the TOEFL Junior® Comprehensive Writing and Speaking Test* (Research Report No. RR-15-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12052
Fan, L., Fan, W., & Huang, D. (2016). Chinese EFL Learners' Public Speaking Skills: An Investigation of Impromptu Speeches at the "FLTRP Cup" English Speaking

Contest in the Light of Communication Models. *Chinese Journal of Applied Linguistics (De Gruyter), 39*, 421–439. https://doi.org/10.1515/cjal-2016-0027
Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT Speaking test for iInternational teaching assistant certification. *Language Assessment Quarterly, 10*, 274–291. https://doi.org/10.1080/15434303.2013.769548
Feng, H.-H. (2014). The use of corpus concordancing for second language learners' self error-correction. *Journal of Interactive Learning Research, 25*(1), 5–25.
Fisher, M. (1985). Rethinking the "Foreign TA Problem.". *New Directions for Teaching and Learning*(22), 63–73. https://doi.org/10.1002/tl.37219852207
Fitch, F., & Morgan, S. E. (2003). 'Not a lick of English': constructing the ITA identity through student narratives. *Communication Education, 52*, 297–310. https://doi.org/10.1080/0363452032000156262
Fox, J., & Cheng, L. (2015). Walk a mile in my shoes: Stakeholder accounts of testing experience with a computer-administered test. *TESL Canada Journal, 32*(9), 65–86. https://doi.org/10.18806/tesl.v32i0.1218
Fox, W. S., & Gay, G. (1994). Functions and Effects of International Teaching Assistants. *Review of Higher Education, 18*, 1–24. https://doi.org/10.1353/rhe.1994.0000
Frase, L. T., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL Test of Written English*™ (Research Report No. RR-98-42). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01791.x
Frase, L. T., Gong, B., Hansen, E. G., Kaplan, R., Katz, I., & Singley, K. (1997). *Technologies for language testing* (Research Memorandum No. RM-97-05). Princeton, NJ: Educational Testing Service.
Freedle, R. O., & Kostin, I. W. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items* (Research Report No. RR-93-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01524.x
Freedle, R. O., & Kostin, I. W. (1996). *The prediction of TOEFL listening comprehension item difficulty for minitalk passages: Implications for construct validity* (Research Report No. RR-96-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1996.tb01707.x
Friedman, M., Sutnick, A. I., Stillman, P. L., Norcini, J. J., Anderson, S. M., Williams, R. G., . . . Reeves, M. J. (1991). The use of standardized patients to evaluate the spoken-English proficiency of foreign medical graduates. *Academic Medicine, 66*(9, Suppl), 61–63. https://doi.org/10.1097/00001888-199109000-00042
Friginal, E., Li, M., & Weigle, S. C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing, 23*, 1–16. https://doi.org/10.1016/j.jslw.2013.10.001
Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning, 21*, 353–367. https://doi.org/10.1080/09588220802343561

Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based test on racial-ethnic and gender groups. *Journal of Educational Measurement, 39*, 133–147. https://doi.org/10.1111/j.1745-3984.2002.tb01139.x

García Gómez, P., Noah, A., Schedl, M., Wright, C., & Yolkut, A. (2007). Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing, 24*, 417–444. https://doi.org/10.1177/0265532207077209

Gardiner, J., & Howlett, S. (2016). Student perceptions of four university gateway tests. *University of Sydney Papers in TESOL, 11*, 67–96.

Getman, E., Cho, Y., & Luce, C. (2016). *Effects of printed option sets on listening item performance among young English-as-a-Foreign-Language learners* (Research Memorandum No. RM-16-16). Princeton, NJ: Educational Testing Service.

Gevara, J. R. (2013). Using Generalizability Theory to examine error variance in the SPEAK scoring rubric. *International Journal of Language Studies, 7*(4), 25–44.

Gholami, J., & Alinasab, M. (2016). Iranian EFL learners' use of self-regulatory, test-wiseness and discourse synthesis strategies in integrated writing tasks. *Pertanika Journal of Social Sciences & Humanities, 24*, 839–853.

Gholami, J., & Alinasab, M. (2017). Source-based tasks in writing independent and integrated essays. *International Journal of Instruction, 10*(3), 127–142. https://doi.org/10.12973/iji.2017.1039a

Ginther, A. (2001). *Effects of the presence and absence of visuals on performance on TOEFL® CBT listening-comprehensive stimuli* (Research Report No. RR-01-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01858.x

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*, 133–167. https://doi.org/10.1191/0265532202lt225oa

Ginther, A., & Elder, C. (2014). *A comparative investigation into understandings and uses of the TOEFL iBT® test, the International English Language Testing Service (Academic) Test, and the Pearson Test of English for graduate admissions in the United States and Australia: A case study of two university contexts* (Research Report No. RR-14-44). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12037

Ginther, A., & Grant, L. (1996). *A review of the academic needs of native English-speaking college students in the United States* (Research Memorandum No. RM-96-04). Princeton, NJ: Educational Testing Service.

Ginther, A., & Yan, X. (2017). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing, 35,* 271–295. https://doi.org/10.1177/0265532217704010

Gleason, J., & Suvorov, R. (2012). Learner Perceptions of Asynchronous Oral Computer-mediated Communication: Proficiency and Second Language Selves. *Canadian Journal of Applied Linguistics, 15*(1), 100–121.

Golub-Smith, M. L. (1987). *A study of the effects of item option rearrangement on the listening comprehension section of the Test of English as a Foreign Langauge* (Research Report No. RR-87-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00221.x

Golub-Smith, M. L., Reese, C., & Steinhaus, K. S. (1993). *Topic and topic comparability on the Test of Written English* (Research Report No. RR-93-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01521.x
González, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal, 8*, 899–919. https://doi.org/10.1002/tesj.342
Goodwin, J., & Gallego, J. C. (1990). The foreign teaching assistant's manual/teaching matters: Skills and strategies for international teaching assistants. *Issues in Applied Linguistics, 1*, 287–293.
Gorsuch, G. (2011). Exporting English pronunciation from China: The communication needs of young Chinese scientists as teachers in higher education abroad. *Forum on Public Policy Online, 2011*(3), 1–27.
Gorsuch, G. (2012). International Teaching Assistants' Experiences in Educational Cultures and Their Teaching Beliefs. *TESL-EJ, 16*, 1–26.
Gorsuch, G. (2016). International teaching assistants at universities: A research agenda. *Language Teaching, 49*, 275–290. https://doi.org/10.1017/S0261444815000452
Gorsuch, G. J. (2003). The Educational Cultures of International Teaching Assistants and U.S. Universities. *TESL-EJ, 7*(3).
Gorsuch, G. J. (2006). Discipline-specific practica for international teaching assistants. *English for Specific Purposes, 25*, 90–108. https://doi.org/10.1016/j.esp.2005.06.003
Gorsuch, G. J. (2011). Improving Speaking Fluency for International Teaching Assistants by Increasing Input. *TESL-EJ, 14*(4).
Gorsuch, G. J., & Sokolowki, J. A. (2007). International Teaching Assistants and Summative and Formative Student Evaluation. *Journal of Faculty Development, 21*(2), 117–136.
Graham, J. C. (1992). Bias-free teaching as a topic in a course for international teaching assistants. *TESOL Quarterly, 26*, 585–589. https://doi.org/10.2307/3587185
Graham, J. G., & Beardsley, R. S. (1986). English for specific purposes: Content, language, and communication in a pharmacy course model. *TESOL Quarterly, 20*, 227–245. https://doi.org/10.2307/3586542
Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31*, 111–133. https://doi.org/10.1177/0265532212469177
Gu, L. (2015). Language ability of young English language learners: Definition, configuration, and implications. *Language Testing, 32*, 21–38. https://doi.org/10.1177/0265532214542670
Gu, L. (2016). Modelling communicative language ability: Skills and contexts of language use. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment* (pp. 17–35). London, England: Bloomsbury Academic.
Gu, L., Lockwood, J., & Powers, D. E. (2015). *Evaluating the TOEFL Junior® Standard Test as a measure of progress for young English language learners* (Research Report No. RR-15–22). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12064
Gu, L., Lockwood, J. R., & Powers, D. E. (2017). Making a validity argument for using the TOEFL Junior standard test as a measure of progress for young English

language learners. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 153–170). New York, NY: Routledge.
Gu, L., & So, Y. (2017). Strategies used by young English learners in an assessment context. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 118–135). New York, NY: Routledge.
Gu, L., & Xi, X. (2015). *Examining performance differences on tests of academic English proficiency used for high-stakes versus practice purposes* (Research Memorandum No. RM-15-09). Princeton, NJ: Educational Testing Service.
Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*, 218–238. https://doi.org/10.1016/j.asw.2013.05.002
Haberman, S. J. (2011). *Use of e-rater® in scoring of the TOEFL iBT® writing test* (Research Report No. RR-11-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02261.x
Haberman, S., & Yao, L. (2015). Repeater analysis for combining information from different assessments. *Journal of Educational Measurement, 52*, 223–251. https://doi.org/10.1111/jedm.12075
Haberman, S. J., Yao, L., & Sinharay, S. (2015). Prediction of true test scores from observed item scores and ancillary data. *British Journal of Mathematical and Statistical Psychology, 68*(2), 363–385. https://doi.org/10.1111/bmsp.12052
Hahn, L. D. (2004). Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals. *TESOL Quarterly, 38*, 201–223. https://doi.org/10.2307/3588378
Hale, G. A. (1988). *The interaction of student major-field group and text content in TOEFL reading comprehension* (Research Report No. RR-88-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00257.x
Hale, G. A. (1992). *Effects of amount of time allowed on the Test of Written English* (Research Report No. RR-92-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01458.x
Hale, G. A., Angelis, P. J., & Thibodeau, L. A. (1980). *Effects of item disclosure on TOEFL performance* (Research Report No. RR-80-34). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1980.tb01231.x
Hale, G. A., & Courtney, R. G. (1991). *Note taking and listening comprehension on the Test of English as a Foreign Language* (Research Report No. RR-91-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01384.x
Hale, G. A., Rock, D. A., & Jirele, T. A. (1989). *Confirmatory factor analysis of the Test of English as a Foreign Language* (Research Report No. RR-89-42). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01327.x
Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. Jr. (1988). *Multiple-choice cloze items and the Test of English as a Foreign Language* (Research Report No. RR-88-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00258.x

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). *Summaries of studies involving the Test of English as a Foreign Language, 1963-1982* (Research Report No. RR-84-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1984.tb00043.x

Hale, G. A., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs* (Research Report No. RR-95-44). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01678.x

Halleck, G. B. (2008). The ITA problem: a ready-to-use simulation. *Simulation & Gaming, 39*(1), 137-146. https://doi.org/10.1177/1046878107308060

Halleck, G. B., & Moder, C. L. (1995). Testing language and teaching skills of international teaching assistants: The limits ofcompensator strategies. *TESOL Quarterly, 29*, 733–758. https://doi.org/10.2307/3588172

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL® 2000—Writing: Composition, community, and assessment* (Research Memorandum No. RM-96-05). Princeton, NJ: Educational Testing Service.

Hamzaoğlu, H., & Koçoğlu, Z. (2016). The application of podcasting as an instructional tool to improve Turkish EFL learners' speaking anxiety. *Educational Media International, 53*, 313–326. https://doi.org/10.1080/09523987.2016.1254889

Hansen, E. G., Forer, D. C., & Lee, M. J. (2004). *Toward accessible computer-based tests: Prototypes for visual and other disabilities* (Research Report No. RR-04-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01952.x

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System, 33*, 107–133. https://doi.org/10.1016/j.system.2004.11.002

Hansen, E. G., & Willut, C. K. (1997). *Computer and communications technologies in colleges and universities of the year 2000* (Research Memorandum No. RM-97-06). Princeton, NJ: Educational Testing Service.

Harsch, C., Ushioda, E., & Ladroue, C. (2017). *Investigating the predictive validity of TOEFL iBT® test scores and their use in informing policy in a United Kingdom university setting* (Research Report No. RR-17-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12167

Hebbani, A., & Hendrix, K. G. (2014). Capturing the Experiences of International Teaching Assistants in the US American Classroom. *New Directions for Teaching & Learning, 2014*(138), 61–72. https://doi.org/10.1002/tl.20097

Hellermann, J., & Vergun, A. (2007). Language which is not taught: The discourse marker use of beginning adult learners of English. *Journal of Pragmatics, 39*, 157–179. https://doi.org/10.1016/j.pragma.2006.04.008

Henning, G. (1990). *A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL listening comprehension item perrformance* (Research Report No. RR-90-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1990.tb01354.x

Henning, G. (1991). *A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items* (Research Report No. RR-91-23).

Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01390.x
Henning, G. (1992). *Scalar analysis of the Test of Written English* (Research Report No. RR-92-30). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01461.x
Henning, G., & Cascallar, E. (1992). *A preliminary study of the nature of communicative competence* (Research Report No. RR-92-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01449.x
Henning, G., Schedl, M. A., & Suomi, B. K. (1995). *Analysis of proposed revisions of the Test of Spoken English* (Research Report No. RR-95-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01636.x
Hertel, T. J., & Sunderman, G. (2009). Student attitudes toward native and non-native language instructors. *Foreign Language Annals, 42*, 468–482. https://doi.org/10.1111/j.1944-9720.2009.01031.x
Hicks, M. M. (1989). *The TOEFL computerized placement test: Adaptive conventional measurement* (Research Report No. RR-89-12). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1989.tb00338.x
Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering, 12*, 145–159. https://doi.org/10.1017/S1351324906004189
Higgins, D., Ramineni, C., & Zechner, K. (2015). The use of learner corpora for automated scoring of written and spoken responses. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 587–586). Cambridge, UK: Cambridge University Press.
Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2010). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language, 25*, 282–306. https://doi.org/10.1016/j.csl.2010.06.001
Hoekje, B., & Linnell, K. (1994). “Authenticity” in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly, 28*, 103–126. https://doi.org/10.2307/3587201
Hsieh, C.-N. (2016). Examining content representativeness of a young learner language assessment: EFL teachers’ perspectives. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 262–284). New York, NY: Springer.
Hsieh, C.-N., Ionescu, M., & Ho, T. (2017). Out of many, one: challenges in teaching multilingual Kenyan primary students in English. *Language, Culture and Curriculum*. Advance online publication. https://doi.org/10.1080/07908318.2017.1378670
Hsieh, C.-N. & Wang, Y. (2017). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532217734240
Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly, 13*, 25–41. https://doi.org/10.1080/15434303.2015.1134540

Huang, H.-T. D., Hung, S.-T. A., & Hong, H.-T. V. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly, 13*, 283–301. https://doi.org/10.1080/15434303.2016.1236111

Huang, H.-T. D., Hung, S.-T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*, 27–49. https://doi.org/10.1177/0265532216677106

Huang, L.-S. (2005). Fine-Tuning the Craft of Teaching by Discussion. *Business Communication Quarterly, 68*, 492–500. https://doi.org/10.1177/108056990506800409

Huang, Y.-P. (2010). International teachers' cross-cultural teaching stories: A tragic comedy. *Curriculum and Teaching Dialogue, 12*, 89–103.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000* (Research Memorandum No. RM-96-06). Princeton, NJ: Educational Testing Service.

In'nami, Y., Koizumi, R., & Nakamura, K. (2016). Factor structure of the Test of English for Academic Purposes (TEAP®) test in relation to the TOEFL iBT® test. *Language Testing in Asia, 6*(1), 1–23. https://doi.org/10.1186/s40468-016-0025-9

Inglis, M. (1993). The communicator style measure applied to nonnative speaking teaching assistants. *International Journal of Intercultural Relations, 17*(1), 89–105. https://doi.org/10.1016/0147-1767(93)90014-Y

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review / La revue canadienne des langues vivantes, 64*, 555–580. https://doi.org/10.3138/cmlr.64.4.555

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*, 401–436. https://doi.org/10.1111/0023-8333.00160

Jackson, R. L., & Hendrix, K. G. (2003). Racial, Cultural, and Gendered Identities in Educational Contexts: Communication Perspectives on Identity Negotiation. *Communication Education, 52*, 177–317.

Jafarpour, A. A., Hashemian, M., & Rafati, M. (2013). The effects of verb tense variation on the fluency in monologues by TOEFL iBT candidates. *Journal of Language Teaching and Research, 4*, 357–362. https://doi.org/10.4304/jltr.4.2.357-362

Jamieson, J., Eignor, D., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 55–95). New York, NY: Routledge.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (Research Memorandum No. RM-00-03). Princeton, NJ: Educational Testing Service.

Jamieson, J., & Poonpon, K. (2013). *Developing analytic rating guides for TOEFL iBT integrated speaking tasks* (Research Report No. RR-13-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02320.x
Jamieson, J., Taylor, C., Kirsch, I. S., & Eignor, D. (1998). Design and evaluation of a computer-based TOEFL tutorial. *System, 26*, 485–513. https://doi.org/10.1016/S0346-251X(98)00034-7
Jamieson, J., Taylor, C., Kirsch, I., & Eignor, D. (1999). *Design and evaluation of a computer-based TOEFL tutorial* (Research Report No. RR-99-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1999.tb01799.x
Jenkins, S. (2000). Cultural and linguistic miscues: A case study of international teaching assistant and academic faculty miscommunication. *International Journal of Intercultural Relations, 24*, 477–501. https://doi.org/10.1016/S0147-1767(00)00011-0
Jenkins, S., & Parra, I. (2003). Multiple Layers of Meaning in an Oral Proficiency Test: The Complementary Roles of Nonverbal, Paralinguistic, and Verbal Behaviors in Assessment Decisions. *Modern Language Journal, 87*, 90–107. https://doi.org/10.1111/1540-4781.00180
Jia, C. L., & Bergerson, A. A. (2008). Understanding the international teaching assistant training program: A case study at a northwestern research university. *International Education, 37*(2), 77–98.
Jiang, X., Sawaki, Y., & Sabatini, J. (2012). Word reading efficiency, text reading fluency, and reading comprehension among Chinese learners of English. *Reading Psychology, 33*, 323–349. https://doi.org/10.1080/02702711.2010.526051
Johncock, P. (1991). International teaching assistants tests and testing policies at U.S. universities. *College & University, 66*, 129–137.
Jung, E. H. (2003). The effects of organization markers on ESL learners' text understanding. *TESOL Quarterly, 37*, 749–759.
Jung, E. H. (2006). Misunderstanding of academic monologues by normative speakers of English. *Journal of Pragmatics, 38*, 1928–1942. https://doi.org/10.1016/j.pragma.2005.05.001
Kalantari, R., & Gholami, J. (2017). Lexical complexity development from dynamic systems theory perspective: Lexical density, diversity, and sophistication. *International Journal of Instruction, 10*(4), 1–18. https://doi.org/10.12973/iji.2017.1041a
Kamiya, N. (2017). Can the National Center Test in Japan be replaced by commercially available private English tests of four skills? In the case of TOEFL Junior Comprehensive. *Language Testing in Asia, 7*(1), 7–15. https://doi.org/10.1186/s40468-017-0046-z
Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*, 301–315. https://doi.org/10.1016/j.system.2010.01.005
Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*, 249–269. https://doi.org/10.1080/15434303.2011.642631

Kang, O., Moran, M., & Thomson, R. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning, 68*, 115–146. https://doi.org/10.1111/lang.12270

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating U.S. Undergraduates' Attitudes Toward International Teaching Assistants. *TESOL Quarterly*, 681–706. https://doi.org/10.1002/tesq.192

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal, 94*, 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O., & Rubin, D. L. (2012). Intra-rater reliability of oral proficiency ratings. International *Journal of Educational and Psychological Assessment, 12*(1), 43–61.

Kaplan, R. B. (1989). The Life and Times of ITA Programs. *English for Specific Purposes, 8*, 109–124. https://doi.org/10.1016/0889-4906(89)90024-0

Katz, I. R., Xi, X., Kim, H.-J., & Cheng, P. C.-H. (2004). *Elicited speech from graph items on the Test of Spoken English* (Research Report No. RR-04-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01933.x

Keck, C. M., & Biber, D. (2004). Modal use in spoken and written university registers. A corpus-based study. In R. Facchinetti & F. Palmer (Eds.), *English modality in perspective: Genre analysis and contrastive studies* (pp. 3–25). Frankfurt am Main, Germany: Peter Lang.

Kim, A.-Y. (2012). Investigating the Effectiveness of Computer-Assisted Language Learning (CALL) in Improving Pronunciation: A Case Study. *Multimedia-Assisted Language Learning, 15*(3), 11–33.

Kim, E.-Y. J. (2017). The TOEFL iBT writing: Korean students' perceptions of the TOEFL iBT writing test. *Assessing Writing, 33*, 1–11. https://doi.org/10.1016/j.asw.2017.02.001

Kim, E. (2009). Beyond language barriers: Teaching self-efficacy among East Asian international teaching assistants. *International Journal of Teaching and Learning in Higher Education, 21*, 171–180.

Kim, H.-J. (2005). Challenge of World Englishes to language testing: Investigation of rater variability in the assessment process. *English Teaching, 60*, 533–548.

Kim, H. J. (2011). Investigating rater behavior across diverse English speaking tasks. *Foreign Languages Education*, *18*(2), 99–125.

Kim, J. (2016). International teaching assistants' initiation of negotiations in engineering labs. *International Journal of Applied Linguistics, 26*, 420–436. https://doi.org/10.1111/ijal.12140

Kim, J. T. (2006). Context validity of a speaking test: Needs analysis and content analysis. *Korean Journal of Applied Linguistics, 22*, 137–158.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*, 89–114. https://doi.org/10.1191/026553201675366418

Kim, S., & Kubota, R. (2012). Supporting Nonnative English-Speaking Instructors to Maximize Student Learning in Their Courses: A Message From the Guest Editors. *Journal on Excellence in College Teaching, 23*(3), 1–6.

Kim, S. H., & Park, I. (2015). Test taker-initiated repairs in an English oral proficiency exam for international teaching assistants. *Text & Talk: An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 237–262. https://doi.org/10.1515/text-2014-0036

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28*, 509–541. https://doi.org/10.1177/0265532211400860

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL examinees* (Research Report No. RR-98-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01755.x

Knoch, U., & Chapelle, C. (2017). Validation of processes within an argument-based framework. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532217710049

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). *An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT® writing test* (Research Report No. RR-14-43). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12038

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing, 29*, 291–308. https://doi.org/10.1177/0265532211429403

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing, 30*, 467–489. https://doi.org/10.1177/0265532213475782

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (Research Report No. RR-04-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01938.x

Kunnan, A., & Carr, N. (2017). A comparability study between the General English Proficiency Test-Advanced and the Internet-Based Test of English as a Foreign Language. *Language Testing in Asia, 7*(1), 1–16. https://doi.org/10.1186/s40468-017-0048-x

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*, 183–189. https://doi.org/10.1177/0265532209349468

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*, 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*, 319–340. https://doi.org/10.1177/0265532215587391

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing, 34*, 451–475. https://doi.org/10.1177/0265532217713951

Lazaraton, A., & Wagner, S. (1996). *The revised Test of Spoken English (TSE): Discourse analysis of native speaker and nonnative speaker data* (Research Memorandum No. RM-96-10). Princeton, NJ: Educational Testing Service.

Leacock, C., & Chodorow, M. (2001). *Automatic assessment of vocabulary usage without negative evidence* (Research Report No. RR-01-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01863.x

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies, 3*(1), 1–134. https://doi.org/10.2200/S00275ED1V01Y201006HLT009

Lee, J. J. (2009). Size matters: An exploratory comparison of small- and large-class university lecture introductions. *English for Specific Purposes, 28*, 42–57. https://doi.org/10.1016/j.esp.2008.11.001

Lee, Y.-W. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks* (Research Memorandum No. RM-04-07). Princeton, NJ: Educational Testing Service.

Lee, Y.-W., Breland, H., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts for different native language groups* (Research Report No. RR-04-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01951.x

Lee, Y.-W., Breland, H., & Muraki, E. (2005). Comparability of TOEFL CBT writing prompts for different native language groups. *International Journal of Testing, 5*, 131–158. https://doi.org/10.1207/s15327574ijt0502_3

Lee, Y.-W., Gentile, C. A., & Kantor, R. (2008). *Analytic scoring of TOEFL® CBT essays: Scores from humans and e-rater®* (Research Report No. RR-08-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02087.x

Lee, Y.-W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics, 31*, 391–417. https://doi.org/10.1093/applin/amp040

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (Research Report No. RR-05-14). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01991.x

Lee, Y.-W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*, 239–263. https://doi.org/10.1080/15434300903079562

Lee, S., & Winke, P. (2017). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing, 35,* 239-269. https://doi.org/10.1177/0265532217704009

LeGros, N., & Faez, F. (2012). The Intersection Between Intercultural Competence and Teaching Behaviors: A Case of International Teaching Assistants. *Journal on Excellence in College Teaching, 23*(3), 7–31.

Li, L., Mazer, J. P., & Ju, R. (2011). Resolving international teaching assistant language inadequacy through dialogue: challenges and opportunities for clarity and

credibility. *Communication Education, 60*, 461–478. https://doi.org/10.1080/03634523.2011.565352
Li, Y., & Brown, T. (2013). *A trend-scoring study for the TOEFL® iBT speaking and writing sections* (Research Memorandum No. RM-13-05). Princeton, NJ: Educational Testing Service.
Lianzhen, H., & Ying, D. (2006). A corpus-based investigation into the validity of the CET–SET group discussion. *Language Testing, 23*, 370–401. https://doi.org/10.1191/0265532206lt333oa
Liao, S. (2009). Variation in the use of discourse markers by Chinese teaching assistants in the US. *Journal of Pragmatics, 41*, 1313–1328. https://doi.org/10.1016/j.pragma.2008.09.026
Ling, G. (2017a). Are TOEFL iBT® writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing, 31*, 1–12. https://doi.org/10.1016/j.asw.2016.04.001
Ling, G. (2017b). Is writing performance related to keyboard type? An investigation from examinees' perspectives on the TOEFL iBT. *Language Assessment Quarterly, 14*, 36–53. https://doi.org/10.1080/15434303.2016.1262376
Ling, G., & Bridgeman, B. (2013). Writing essays on a laptop or a desktop computer: Does it matter? *International Journal of Testing, 13*, 105–122. https://doi.org/10.1080/15305058.2012.690012
Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*, 479–499. https://doi.org/10.1177/0265532214530699
Ling, G., Powers, D. E., & Adler, R. M. (2014). *Do TOEFL iBT® scores reflect improvement in English-language proficiency? Extending the TOEFL iBT validity argument* (Research Report No. RR-14-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12007
Ling, G., Wolf, M. K., Cho, Y., & Wang, Y. (2014). *English-as-a-second-language programs for matriculated students in the United States: An exploratory survey and some issues* (Research Report No. RR-14-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12010
Liu, O. L. (2011). Do major field of study and cultural familiarity affect TOEFL iBT reading performance? A confirmatory approach to differential item functioning. *Applied Measurement in Education, 24*, 235–255. https://doi.org/10.1080/08957347.2011.580645
Liu, O. L. (2014). *Investigating the relationship between test preparation and TOEFL® iBT performance* (Research Report No. RR-14-15). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12016
Liu, O. L., Schedl, M. A., Malloy, J., & Kong, N. (2009). *Does content knowledge affect TOEFL iBT™ reading performance? A confirmatory approach to differential item functioning* (Research Report No. RR-09-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02186.x
Llosa, L., & Malone, M. E. (2017). Student and instructor perceptions of writing tasks and performance on TOEFL iBT versus university writing courses. *Assessing Writing, 34*, 88–99. https://doi.org/10.1016/j.asw.2017.09.004

LoCastro, V., & Tapper, G. (2006). International Teaching Assistants and teacher identity. *Journal of Applied Linguistics, 3*, 185–218. https://doi.org/10.1558/japl.v3i2.185
Longford, N. T. (1996). *Adjustment for reader rating behavior in the Test of Written English* (Research Report No. RR-95-39). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01673.x
Looney, S. D., Jia, D., & Kimura, D. (2017). Self-directed okay in mathematics lectures. *Journal of Pragmatics, 107*, 46–59. https://doi.org/10.1016/j.pragma.2016.11.007
Loukina, A., & Buzick, H. (2017). *Use of automated scoring in spoken language assessments for test takers with speech impairments* (Research Report No. RR-17-42). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12170
Ma, J., & Cheng, L. (2015). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth, and significance. *TESL Canada Journal, 33*(1), 58–79. https://doi.org/10.18806/tesl.v33i1.1227
Ma, R. (2014). English Communication for International Teaching Assistants. *Journal of International Students*, *4,* 199–201.
MacMillan, F. (2007). The role of lexical cohesion in the assessment of EFL reading proficiency. *Arizona Working Papers in SLA and Teaching, 14*, 75–93.
Magno, C., de Carvalho Filho, M. K., & Lajom, J. A. (2011). Factors Involved in the Use of Language Learning Strategies and Oral Proficiency Among Taiwanese Students in Taiwan and in the Philippines. *Asia-Pacific Education Researcher (De La Salle University Manila), 20*, 489–502.
Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balastubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*, 173–190. https://doi.org/10.2307/3588329
Malone, M. E., & Montee, M. (2014). *Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability* (Research Report No. RR-14-42). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12039
Manna, V. F., & Yoo, H. (2015). *Investigating the relationship between test-taker background characteristics and test performance in a heterogeneous English-as-a-Second-Language (ESL) test population: A factor analytic approach* (Research Report No. RR-15-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12072
Manning, W. H. (1987). *Development of cloze-elide tests of English as a second language* (Research Report No. RR-87-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00222.x
Manohar, U., & Appiah, O. (2016). Perspective Taking to Improve Attitudes towards International Teaching Assistants: The Role of National Identification and Prior Attitudes. *Communication Education, 65*, 149–163. https://doi.org/10.1080/03634523.2015.1081956
Marasco, J. (1993). Communicate: Strategies for International Teaching Assistants by Jan Smith Colleen M. Meyers Amy J. Burkhalter. *TESOL Quarterly*, *27,* 773–774. https://doi.org/10.2307/3587424
McCrocklin, S. (2011). [English Communication for International Teaching Assistants]. *TESL-EJ, 14*(4), F1–F3.

McCullough, A. S. (1996). Discourse and Performance of International Teaching Assistants. *Studies in Second Language Acquisition, 18*, 517–518. https://doi.org/10.1017/S027226310001545X

McKinley, R. L., & Way, W. D. (1992). *The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models* (Research Report No. RR-92-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01448.x

Mehra, B., & Papajohn, D. (2007). "Glocal" patterns of communication-information convergences in Internet use: Cross-cultural behavior of international teaching assistants in a culturally alien information environment. *International Information and Library Review, 39*, 12–30. https://doi.org/10.1016/j.iilr.2007.01.001

Meyer, K. R., & Mao, Y. (2014). Comparing Student Perceptions of the Classroom Climate Created by U.S. American and International Teaching Assistants. *Higher Learning Research Communications, 4*(3), 12–22. https://doi.org/10.18870/hlrc.v4i3.206

Minchew, S. S., & Couvillion, M. B. (2010). A COMPARISON OF AMERICAN AND INTERNATIONAL STUDENTS' LIFETSYLES AND PERCEPTIONS OF THE UNIVERSITY EXPERIENCE. *National Forum of Applied Educational Research Journal, 23*(3), 1–8.

Mislevy, R. J., & Yin, C. (2012). Evidence-Centered Design in Language Testing. In G. Fulcher & F. Davidson (Eds.), The Routledge Handbook of Language Testing. London: Routledge.

Molholt, C. (1988). Computer-Assisted Instruction in Pronunciation for Chinese Speakers of American English. *TESOL Quarterly, 22*, 91–111. https://doi.org/10.2307/3587063

Monoson, P. K., & Thomas, C. F. (1993). Oral English Proficiency Policies for Faculty in U.S. Higher Education. *Review of Higher Education, 16*, 127–140. https://doi.org/10.1353/rhe.1993.0020

Mousavi, S. A. (2009). Multimedia as a test method facet in oral proficiency tests. *International Journal of Pedagogies & Learning, 5*, 37–48. https://doi.org/10.5172/ijpl.5.1.37

Mustafa, F., & Apriadi, H. (2016). Diy: Designing a reading test as reliable as a paper-based TOEFL designed by ETS. *Proceedings of English Education International Conference, 1*(2), 402–407.

Myers, S. A. (1995). Using written text to teach oral skills: An ITA training class using field-specific materials. *English for Specific Purposes, 14*, 231–245. https://doi.org/10.1016/0889-4906(95)00011-1

Myford, C. M., Marr, D., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English* (Research Report No. RR-95-40). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01674.x

Myford, C. M., & Wolfe, E. W. (2000a). *Monitoring sources of variability within the Test of Spoken English assessment system* (Research Report No. RR-00-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01829.x

Myford, C. M., & Wolfe, E. W. (2000b). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (Research Report No. RR-00-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01832.x

Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement, 3*, 300–324.

Nahal, A. (2005). Cultural collisions: the relationship between international teaching assistants and American students can often be challenging--tips to ease the tension. *Diverse Issues in Higher Education, 22*(20), 41.

Nelson, G. L. (1992). The relationship between the use of personal, cultural examples in international teaching assistants' lectures and uncertainty reduction, student attitude, student recall, and ethnocentrism. *International Journal of Intercultural Relations, 16*, 33–52. https://doi.org/10.1016/0147-1767(92)90004-E

Nguyen, B. B.-D. (1993). Accent discrimination and the Test of Spoken English: A call for an objective assessment of the comprehensibility of nonnative speakers. *California Law Review, 81*(5), 1325–1361. https://doi.org/10.2307/3480920

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension* (Research Report No. RR-95-37). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01671.x

Nissan, S., & Schedl, M. A. (2012). Prototyping New Item Types. In G. Fulcher & F. Davidson (Eds.), The Routledge Handbook of Language Testing. London: Routledge.

Nofsinger, R. E. (1990). Measuring minimal proficiency in spoken English. *Communication Reports, 3*, 37–44. https://doi.org/10.1080/08934219009367499

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics, 37*, 693–715. https://doi.org/10.1093/applin/amu060

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing, 32*, 39–62. https://doi.org/10.1177/0265532214538014

Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of strength of accent on an L2 interactive lecture listening comprehension test. *International Journal of Listening, 30*(1–2), 84–98. https://doi.org/10.1080/10904018.2015.1056877

Olaniran, B. A. (1999). International Graduate Teaching Assistants Workshop: Implications for Training. *College Student Affairs Journal, 18*(2), 56–71.

Oltman, P. K., Stricker, L. J., & Barrows, T. S. (1988). *Native language, English proficiency, and the structure of the Test of English as a Foreign Language* (Research Report No. RR-88-26). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00282.x

Oltman, P. K., & Stricker, L. J. (1989). *Developing homogeneous TOEFL scales by multidimensional scaling* (Research Report No. RR-89-44). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01329.x

Papageorgiou, S., & Baron, P. A. (2017). Using the Common European Framework of Reference for young learners' English language proficiency assessments. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 136–152). New York, NY: Routledge.

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL® Junior™ Standard scores for ESL placement decisions in secondary education. *Language Testing, 31*, 223–239. https://doi.org/10.1177/0265532213499750

Papageorgiou, S., Morgan, R., & Becker, V. (2015). Enhancing the Interpretability of the overall results of an international test of English-language proficiency. *International Journal of Testing, 15*, 310–336. https://doi.org/10.1080/15305058.2015.1078335

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly, 12*, 153–177. https://doi.org/10.1080/15434303.2015.1008480

Papajohn, D. (1999). The Effect of Topic Variation in Performance Testing: The Case of the Chemistry TEACH Test for International Teaching Assistants. *Language Testing, 16*, 52–81. https://doi.org/10.1177/026553229901600104

Park, S.Y. (2013). Korean EFL learners' use of cohesive devices in narrative and argumentative essays. *Studies in English Education, 18*(2), 51–81.

Peirce, B. N. (1992). Demystifying the TOEFL® reading test. *TESOL Quarterly, 26*, 665–691. https://doi.org/10.2307/3586868

Perlmutter, M. (1989). Intelligibility rating of L2 speech pre- and postintervention. *Perceptual and Motor Skills*, 515–521. https://doi.org/10.2466/pms.1989.68.2.515

Pickering, L. (2001). The Role of Tone Choice in Improving ITA Communication in the Classroom. *TESOL Quarterly, 35*, 233–255. https://doi.org/10.2307/3587647

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes, 23*, 19–43. https://doi.org/10.1016/S0889-4906(03)00020-6

Pike, L. W. (1979). *An evaluation of alternative item formats for testing English as a foreign language* (Research Report No. RR-79-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1979.tb01174.x

Plakans, B. S. (1997). Undergraduates' Experiences with and Attitudes toward International Teaching Assistants. *TESOL Quarterly, 31*, 95–119. https://doi.org/10.2307/3587976

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*, 217–230. https://doi.org/10.1016/j.jslw.2013.02.003

Plakans, L. & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing*, *31*, 98–112. https://doi.org/10.1016/j.asw.2016.08.005

Plakans, L., Gebril, A., & Bilki, Z. (2016). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532216669537

Powell, W. W. (2001). *Looking back, looking forward: Trends in intensive English program enrollments* (Research Memorandum No. RM-01-01). Princeton, NJ: Educational Testing Service.

Powers, D. E. (1980). *The relationship between scores on the Graduate Management Admission Test and the Test of English as a Foreign Language* (Research Report No. RR-80-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1980.tb01228.x

Powers, D. E. (1985). *A survey of academic demands related to listening skills* (Research Report No. RR-85-48). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1985.tb00133.x

Powers, D. E. (2011). *Scoring the TOEFL® independent essay automatically: Reactions of test takers and test score users* (Research Memorandum No. RM-11-34). Princeton, NJ: Educational Testing Service.

Powers, D. E., Albertson, W., Florek, T., Johnson, K., Malak, J., Nemceff, B., …, Zelazny, A. (2002). *Influence of irrelevant speech on standardized test performance* (Research Report No. RR-02-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2002.tb01873.x

Powers, D. E., & Lall, V. F. (2012). *Supporting an expiration policy for TOEFL® scores* (Research Memorandum No. RM-12-03). Princeton, NJ: Educational Testing Service.

Powers, D. E., Roever, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge™ Courseware against faculty ratings and student self-assessments* (Research Report No. RR-03-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01903.x

Powers, D., Schedl, M., & Papageorgiou, S. (2017). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing, 34*, 175–195. https://doi.org/10.1177/0265532215623582

Powers, D. E., Schedl, M. A., Wilson-Leung, S., & Butler, F. A. (1999a). *Validating the revised Test of Spoken English against a criterion of communicative success* (Research Report No. RR-99-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1999.tb01803.x

Powers, D. E., Schedl, M. A., Wilson-Leung, S., & Butler, F. A. (1999b). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*, 399–425. https://doi.org/10.1177/026553229901600401

Powers, D. E., & Stansfield, C. W. (1983). *The Test of Spoken English as a measure of communicative ability in the health professions: Validation and standard setting* (Research Report No. RR-83-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1983.tb00001.x

Powers, D. E., & Stansfield, C. W. (1985). Testing the oral English proficiency of foreign nursing graduates. *The ESP Journal, 4*(1), 21–35. https://doi.org/10.1016/0272-2380(85)90004-6

Powers, D. E., & Stansfield, C. W. (1989). An Approach to the Measurement of Communicative Ability in Three Health Professions. In H. Coleman (Ed.), Working with Language: A Multidisciplinary Consideration of Language Use in Work Contexts (pp. 341-366). Berlin: Mouton de Gruyter.

Quiñones, A. C., & Humphrey, G. E. (2007). Addressing the pharmacist shortage through a cooperative internship program for foreign pharmacy graduates. *Journal of the American Pharmacists Association, 47*, 191–196. https://doi.org/10.1331/QJ0N-4278-738U-0061

Rahimi, F., Bagheri, M. S., Sadighi, F., & Yarmohammadi, L. (2014). Using an argument-based approach to ensure fairness of high-stakes tests' score-based consequences. *Procedia - Social and Behavioral Sciences, 98*, 1461–1468. https://doi.org/10.1016/j.sbspro.2014.03.566

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the TOEFL® independent and integrated prompts* (Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02288.x

Rasskazova, T., Muzafarova, A., Daminova, J., & Okhotnikova, A. (2017). Computerised language assessment: Limitations and opportunites. *eLearning & Software for Education, 2*, 173–180. https://doi.org/10.12753/2066-026X-17-110

Reed, D. J., & Bowles, M. (2008). A Life in Language Testing: An Interview with Charles Stansfield. *Language Assessment Quarterly*, *5*, 336–359.

Reinhardt, J. (2010). Directives in office hour consultations: A corpus-informed investigation of learner and expert usage. *English for Specific Purposes, 29*, 94–107. https://doi.org/10.1016/j.esp.2009.09.003

Reinhardt, J. (2013). An applied genre analysis of office hours consultations. *International Journal of Corpus Linguistics, 18*, 301–326. https://doi.org/10.1075/ijcl.18.3.03rei

Rekhart, D., & Dunkel, P. (1992). The utility of objective (computer) measures of the fluency of speakers of English as a second language. *Applied Language Learning, 3*(1–2), 65–85.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition, 23*, 497–526. https://doi.org/10.1017/S027226310100403X

Riazi, A. M. (2016). Comparing writing performance in TOEFL iBT and academic assignments: An exploration of textual features. *Assessing Writing, 28*, 15–27. https://doi.org/10.1016/j.asw.2016.02.001

Roach, K. D., & Olaniran, B. A. (2001). Intercultural willingness to communicate and communication anxiety in international teaching assistants. *Communication Research Reports, 18*, 26–35. https://doi.org/10.1080/08824090109384779

Roever, C., & Powers, D. E. (2005). *Effects of language of administration on a self-assessment of language skills* (Research Memorandum No. RM-04-06). Princeton, NJ: Educational Testing Service.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (Research Memorandum No. RM-01-03). Princeton, NJ: Educational Testing Service.

Rosenfeld, M., Oltman, P. K., & Sheppard, K. (2004). *Investigating the validity of TOEFL: A feasibility study using content and criterion-related strategies* (Research Report No. RR-03-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01910.x

Ross, P. G., & Krider, D. S. (1992). Off the plane and into the classroom: A phenomenological explication of international teaching assistants' experiences in the American classroom. *International Journal of Intercultural Relations, 16*, 277–293. https://doi.org/10.1016/0147-1767(92)90053-W

Rounds, P. L. (1987). Multifunctional personal pronoun use in an educational setting. *English for Specific Purposes, 6*, 13–29. https://doi.org/10.1016/0889-4906(87)90072-X

Rubin, D. L. (1993). The other half of international teaching assistant training: Classroom communication workshops for international students. *Innovative Higher Education, 17*, 183–193. https://doi.org/10.1007/BF00915600

Saif, S. (2006). Aiming for positive washback: a case study of international teaching assistants. *Language Testing, 23*, 1–34. https://doi.org/10.1191/0265532206lt322oa

Salomone, A. M. (1998). Communicative Grammar Teaching: A Problem for and a Message from International Teaching Assistants. *Foreign Language Annals, 31*, 552–566. https://doi.org/10.1111/j.1944-9720.1998.tb00599.x

Sarkodie-Mensah, K. (1991). The International Student as TA: A Beat from a Foreign Drummer. *College Teaching, 39*, 115–116. https://doi.org/10.1080/87567555.1991.10532442

Sarwark, S. M., Smith, J., MacCallum, R., & Cascallar, E. C. (1995). *A study of characteristics of the SPEAK test* (Research Report No. RR-94-47). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01620.x

Sawaki, Y. (2017). *The effects of different levels of performance feedback on TOEFL iBT® reading practice test performance* (Research Report No. RR-17-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12159

Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*, 190–209. https://doi.org/10.1080/15434300902801917

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section* (Research Report No. RR-09-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02159.x

Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly, 10*, 73–95. https://doi.org/10.1080/15434303.2011.633305

Sawaki, Y., & Sinharay, S. (2013). *Investigating the value of section scores for the TOEFL iBT® Test* (Research Report No. RR-13-35). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02342.x

Sawaki, Y., & Sinharay, S. (2017). Do the TOEFL iBT section scores provide value-added information to stakeholders? *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532217716731

Sawaki, Y., Stricker, L. J., & Oranje, A. (2008). *Factor structure of the TOEFL® Internet-based Test (iBT): Exploration in a field trial sample* (Research Report No. RR-08-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02095.x

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*, 5–30. https://doi.org/10.1177/0265532208097335

Schedl, M. A., Gordon, A., Carey, P. A., & Tang, K. L. (1996). *An analysis of the dimensionality of TOEFL reading comprehension items* (Research Report No. RR-95-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01662.x

Schedl, M. A., Thomas, N., & Way, W. D. (1994). *An investigation of proposed revisions to section 3 of the TOEFL test* (Research Report No. RR-94-42). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01615.x

Schmidgall, J., Getman, E., & Zu, J. (2017). Screener tests need validation too: Weighing an argument for test use against practical concerns. *Language Testing*. Advance online publication. https://doi.org/10.1177/0265532217718600

Schneider, M., & Connor, U. (1990). Analyzing topical structure in ESL essays: Not all topics are equal. *Studies in Second Language Acquisition, 12*, 411–427. https://doi.org/10.1017/S0272263100009505

Secolsky, C. (1989). *Accounting for random responding at the end of the test in assessing speededness* (Research Report No. RR-89-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1989.tb00337.x

Sequeira, D.-L., & Costantino, M. (1989). Issues in ITA Training Programs. *New Directions for Teaching and Learning,* (39), 79–86. https://doi.org/10.1002/tl.37219893909

Shahrzad, S. (2002). A Needs-Based Approach to the Evaluation of the Spoken Language Ability of International Teaching Assistants. *Canadian Journal of Applied Linguistics*, 145–167.

Shamsaee, S., & Zahedi, K. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): evidence from Iranian test-takers. *Educational Assessment Evaluation and Accountability, 24*, 263–277. https://doi.org/10.1007/s11092-011-9137-z

Sheehan, K. M. (2017). *Helping students select appropriately challenging text: Application to a test of second language reading ability* (Research Report No. RR-17-33). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12160

Shermis, M., Burstein, J., Brew, C., Higgins. D., & Zechner, K. (2015). Recent innovations in machine scoring of student and test taker written and spoken responses. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (pp. 335–354). New York, NY: Routledge.

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Peterson, E. Baker, & B. McGaw

(Eds.), *International encyclopedia of education* (3rd ed., pp. 20–26). Oxford, UK: Elsevier.
Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*, 31–57. https://doi.org/10.1191/0265532205lt296oa
Shirzad, M., & Shirzad, H. (2017). The effect of computer literacy on the participants' writing ability in TOEFL iBT. *Theory & Practice in Language Studies, 7*, 134–139. https://doi.org/10.17507/tpls.0702.07
Smith, J. (1989). Topic and variation in ITA oral proficiency: SPEAK and field-specific tests. *English for Specific Purposes, 8*, 155–167. https://doi.org/10.1016/0889-4906(89)90027-6
Smith, K. S., & Simpson, R. D. (1993). Becoming Successful as an International Teaching Assistant. *Review of Higher Education, 16*, 483–497. https://doi.org/10.1353/rhe.1993.0009
Smith, R. A., Strom, R. E., & Muthuswamy, N. (2005). Undergraduates' Ratings of Domestic and International Teaching Assistants: Timing of Data Collection and Communication Intervention. *Journal of Intercultural Communication Research, 34*(1/2), 3–21.
Smith, R. M., Byrd, P., Nelson, G., Barrett, R. P., & Constantinides, J. C. (1995). Crossing pedagogical oceans - International teaching assistants in the US undergraduate education. *Modern Language Journal, 79*, 122–123.
Smith, R. M., Byrd, P., & Nelson, G. L. (1992). Crossing pedagogical oceans: international teaching assistants in U.S. undergraduate education. *ASHE-ERIC Higher Education Report*(8), 1–110.
So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly, 11*, 283–303. https://doi.org/10.1080/15434303.2014.936936
So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior® design framework* (Research Report No. RR-15-13). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12058
Southard, B., & Sheorey, R. (1992). Measures of oral proficiency in English; A comparison of the rated interview and the Test of Spoken English. *College ESL, 2*(2), 52–67.
Stankov, L., & Lee, J. (2008). Confidence and Cognitive Test Performance. *Journal of Educational Psychology, 100*, 961–976. https://doi.org/10.1037/a0012546
Stankov, L., Lee, J., & Paek, I. (2009). Realism of confidence judgments. *European Journal of Psychological Assessment, 25*, 123–130. https://doi.org/10.1027/1015-5759.25.2.123
Stansfield, C. W. (1985). *Toward communicative competence testing: Proceedings of the Second TOEFL® invitational conference* (TOEFL Research Report No. 21). Princeton, NJ: Educational Testing Service.
Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the Test of Written English. *Language Testing, 5*, 160–186. https://doi.org/10.1177/026553228800500204

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*, 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Staples, S., Kang, O., & Wittner, E. (2014). Considering interlocutors in university discourse communities: Impacting US undergraduates' perceptions of ITAs through a structured contact program. *English for Specific Purposes, 35*, 54–65. https://doi.org/10.1016/j.esp.2014.02.002

Stenson, N., Downing, B., Smith, J., & Smith, K. (1992). The Effectiveness of Computer-Assisted Pronunciation Training. *CALICO Journal, 9*(4), 5–19.

Stevens, S. G. (1989). A "dramatic" approach to improving the intelligibility of ITAs. *English for Specific Purposes, 8*, 181–194. https://doi.org/10.1016/0889-4906(89)90029-X

Stevenson, I., & Jenkins, S. (1994). Journal writing in the training of international teaching assistants. *Journal of Second Language Writing, 3*, 97–120. https://doi.org/10.1016/1060-3743(94)90010-8

Storch, N., & Hill, K. (2008). What happens to international students' English after one semester at university? *Australian Review of Applied Linguistics, 31*(1), 4.1-4.17. https://doi.org/10.2104/aral0804

Stricker, L. J. (1997). *Using just noticeable differences to interpret Test of Spoken English scores* (Research Report No. RR-97-04). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02095.x

Stricker, L. J. (2000). Using just noticeable differences to interpret test scores. *Psychological Methods, 5*(4), 415–424. https://doi.org/10.1037/1082-989X.5.4.415

Stricker, L. J. (2002). *The performance of native speakers of English and ESL speakers on the TOEFL® CBT and GRE® General Test* (Research Report No. RR-02-16). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2002.tb01883.x

Stricker, L. J. (2004). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. *Language Testing, 21*, 146–173. https://doi.org/10.1191/0265532204lt279oa

Stricker, L. J., & Attali, Y. (2010). *Test takers' attitudes about the TOEFL iBT*™ (Research Report No. RR-10-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02209.x

Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL Internet-Based Test across subgroups* (Research Report No. RR-08-66). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02152.x

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge™ Test across language groups* (Research Report No. RR-05-12). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01989.x

Stricker, L. J., & Wilder, G. Z. (2001). *Examinees' attitudes about the TOEFL CBT, possible determinants, and relationships with test performance* (Research Report No. RR-01-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01843.x

Stricker, L. J., & Wilder, G. Z. (2012). *Test takers' interpretation and use of TOEFL iBT score reports: A focus study group* (Research Memorandum No. RM-12-08). Princeton, NJ: Educational Testing Service.

Subtirelu, N. C. (2015). "She does have an accent but ...": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society, 44*, 35–62. https://doi.org/10.1017/S0047404514000736

Subtirelu, N. C. (2017). Students' orientations to communication across linguistic difference with international teaching assistants at an internationalizing university in the United States. *Multilingua: Journal of Cross-Cultural and Interlanguage Communication, 36*, 247–280. https://doi.org/10.1515/multi-2016-0061

Sungatullina, D., D., Zalyaeva, E., O., & Gorelova, Y., N. (2016). Metacognitive awareness of TOEFL reading comprehension strategies. *Mediterranian Journal of Social Sciences, 6*(6), 430–436. https://doi.org/10.1051/shsconf/20162601046

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT™ (SSTiBT): Test-takers' reported strategic behaviors* (Research Report No. RR-09-30). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02187.x

Swinton, S. S. (1983). *A manual for assessing language growth in instructional settings* (Research Report No. RR-83-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1983.tb00017.x

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis of the Test of English as a Foreign Language for several language groups* (Research Report No. RR-80-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1980.tb01229.x

Tang, K. L. (1996). *Polytomous item response theory (IRT) models and their applications in large-scale testing programs: Review of literature* (Research Memorandum No. RM-96-08). Princeton, NJ: Educational Testing Service.

Tang, K. L., & Eignor, D. R. (1997). *Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models* (Research Report No. RR-97-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1997.tb01727.x

Tang, K. L., & Eignor, D. R. (2001). *A study of the use of collateral statistical information in attempting to reduce TOEFL IRT item parameter estimation sample sizes* (Research Report No. RR-01-11). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01853.x

Tang, K. L., Way, W. D., Carey, P. A. (1993). *The effect of small calibration sample sizes on TOEFL IRT-based equating* (Research Report No. RR-93-59). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01570.x

Tannenbaum, R. J., & Baron, P. A. (2011). *Mapping TOEFL ITP scores onto the Common European Framework of Reference* (Research Memorandum No. RM-11-33). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping scores from the TOEFL Junior Comprehensive Test onto the Common European Framework of Reference (CEFR)* (Research Memorandum No. RM-15-13). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J., & Wylie, E. C. (2005). *Mapping English language proficiency test scores onto the Common European Framework* (Research Report No. RR-05-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01995.x

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology* (Research Report No. RR-08-34). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (Research Report No. RR-98-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01757.x

Thomas, C. F., & Monoson, P. K. (1993). Oral English language proficiency of ITAs: Policy, implementation, and contributing factors. *Innovative Higher Education, 17*, 195–209. https://doi.org/10.1007/BF00915601

Timpe-Laughlin, V. (2015). *Evaluating a learning tool for youngl English learners: The case of the TOEFL Primary English Learning Center* (Research Memorandum No. RM-15-04). Princeton, NJ: Educational Testing Service.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*, 713–729. https://doi.org/10.2307/3586870

Tyler, A., & Boxer, D. (1996). Sexual harassment? Cross-cultural/cross-linguistic perspectives. *Discourse & Society*, 107–133. https://doi.org/10.1177/0957926596007001005

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

Wagner, E. (2016). *A study of the use of the TOEFL iBT® test speaking and listening scores for international teaching assistant screening* (Research Report No. RR-16-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12104

Wainer, H., & Lukhele, R. (1997). *How reliable is the TOEFL test?*(Research Report No. RR-97-08). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1997.tb01729.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203–220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

Wainer, H., & Wang, X. (2001). *Using a new statistical model for testlets to score TOEFL* (Research Report No. RR-01-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01851.x

Wall, D., & Horák, T. (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline*

study (Research Report No. RR-06-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02024.x

Wall, D., & Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education: Principles, Policy & Practice, 14*, 99–116. https://doi.org/10.1080/09695940701272922

Wall, D., & Horák, T. (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change* (Research Report No. RR-08-37). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02123.x

Wall, D., & Horák, T. (2011). *The impact of changes in the TOEFL exam on teaching in a sample of countries in Europe: Phase 3, the role of the coursebook. Phase 4, Describing change* (Research Report No. RR-11-41). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02277.x

Wallace, L. (2015). Reflexive photography, attitudes, behavior, and CALL: ITAs improving spoken English intelligibility. *CALICO Journal, 32*, 449–479. https://doi.org/10.1558/cj.v32i3.26384

Wang, B. (2008). What can we tell from these temporal measures?--Temporal measures as indices of oral proficiency. *English Language Teaching, 1*(2), 21–31. https://doi.org/10.5539/elt.v1n2p21

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*, 109–128.

Wang, Z., Zechner, K., & Sun, Y. (2016). Monitoring the performance of human and automated scores for spoken responses. *Language Testing, 35,* 101-120. https://doi.org/10.1177/0265532216679451

Waters, A. (1996). *A review of research into needs in English for Academic Purposes of relevance to the North American higher education context* (Research Memorandum No. RM-96-07). Princeton, NJ: Educational Testing Service.

Way, W. D., Carey, P. A., & Golub-Smith, M. L. (1992). *An exploratory study of characteristics related to IRT item parameter invariance with the Test of English as a Foreign Language* (Research Report No. RR-92-43). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01474.x

Way, W. D., & McKinley, R. L. (1991). *Development of procedures for resolving irregularities in the administration of the listening comprehension section of the TOEFL test* (Research Report No. RR-91-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1991.tb01372.x

Way, W. D., & Reese, C. M. (1990). *An investigation of the use of simplified IRT models for scaling and equating the TOEFL test* (Research Report No. RR-90-29). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1990.tb01365.x

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly, 12*, 283–304. https://doi.org/10.1080/15434303.2015.1037446

Wei, M. (2011). A comparative study of the oral proficiency of Chinese learners of English across task functions: A discourse marker perspective. *Foreign Language Annals, 44*, 674–691. https://doi.org/10.1111/j.1944-9720.2011.01156.x

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*, 335–353. https://doi.org/10.1177/0265532210364406

Weigle, S. C. (2011). *Validation of automated scores of TOEFL iBT tasks against nontest indicators of writing ability* (Research Report No. RR-11-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02260.x

Weimer, M., Svinicki, M. D., & Bauer, G. (1989). Designing Programs to Prepare TAs to Teach. *New Directions for Teaching and Learning,* (39), 57–70. https://doi.org/10.1002/tl.37219893907

Wendt, A., & Woo, A. (2009). A Minimum English Proficiency Standard for the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT). *National Council of State Boards of Nursing, 86*(19), 1–10.

Williams, G. M. (2011). Examining classroom negotiation strategies of international teaching assistants. *International Journal for the Scholarship of Teaching & Learning, 5*(1), 1–16. https://doi.org/10.20429/ijsotl.2011.050121

Williams, G. M., & Case, R. E. (2015). Tale of the tape: International teaching assistant noticing during videotaped classroom observations. *Journal of International Students, 5*, 434–446.

Williams, J. (1992). Planning, discourse marking and the comprehensibility of international teaching assistants. *TESOL Quarterly, 26*, 693–711. https://doi.org/10.2307/3586869

Williamson, D. M., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement, Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, K. M. (1982). *GMAT and GRE aptitude test performance in relation to primary language and scores on TOEFL* (Research Report No. RR-82-28). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01314.x

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language* (Research Report No. RR-87-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00207.x

Wilson, K. M. (1993). *Uses of the Secondary Level English Proficiency (SLEP®) test: A survey of current practice* (Research Report No. RR-93-09). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1993.tb01520.x

Wilson, K. M., & Harrison, R. H. (1982). *A comparative analysis of TOEFL examinee characteristics, 1977-1979* (Research Report No. RR-82-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01313.x

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*, 762–789. https://doi.org/10.1002/tesq.73

Winke, P., Gass, S., & Myford, C. (2011). *The relationship between raters' prior language study and the evaluation of foreign language speech samples* (Research Report No. RR-11-30). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02266.x

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*, 231–252. https://doi.org/10.1177/0265532212456968

Wolf, M. K., & Butler, Y. G. (2017). An overview of English language proficiency assessments for young learners. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 3–21). New York, NY: Routledge.

Wolf, M. K., & Steinberg, J. (2011). *An examination of United States middle school students' performance on TOEFL Junior* (Research Memorandum No. RM-11-15). Princeton, NJ: Educational Testing Service.

Wolfe, E. W. (2003). Examinee characteristics associated with choice of composition medium on the TOEFL writing section. *The Journal of Technology, Learning, and Assessment, 2*(4), 1–25.

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology, 8*(1), 53–65. https://doi.org/hdl.handle.net/10125/25229

Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL writing scores* (Research Report No. RR-04-29). Princeton, NJ: Educational Testing Service.

Wu, J., & Lee, M. (2017). The relationships between test performance and students' perceptions of learning motivation, test value, and test anxiety in the context of the English benchmark requirement for graduation in Taiwan's universities. *Language Testing in Asia, 7*(1), 1–21. https://doi.org/10.1186/s40468-017-0041-4

Wylie, E. C., & Tannenbaum, R. J. (2006). *TOEFL academic speaking test: Setting a cut score for international teaching assistants* (Research Memorandum No. RM-06-01). Princeton, NJ: Educational Testing Service.

Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing, 22*, 463–508. https://doi.org/10.1191/0265532205lt305oa

Xi, X. (2007a). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24*, 251–286. https://doi.org/10.1177/0265532207076365

Xi, X. (2007b). Validating TOEFL® iBT speaking and setting score requirements for ITA screening. *Language Assessment Quarterly, 4*, 318–351. https://doi.org/10.1080/15434300701462796

Xi, X. (2008a). Erratum for Xi (2007). *Language Assessment Quarterly, 5*, 87. https://doi.org/10.1080/15434300801893627

Xi, X. (2008b). *Investigating the criterion-related validity of the TOEFL® speaking scores for ITA screening and setting standards for ITAs* (Research Report No. RR-08-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02088.x

Xi, X. (2008c). What and how much evidence do we need? Critical considerations in validating an automated scoring system. In C. A. Chapelle, Y.-R. Chung & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment: Selected papers from the fifth annual conference on*

technology for second language learning (pp. 102–114). Ames, IA: Iowa State University.
Xi, X. (2010a). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing, 27*, 73–100. https://doi.org/10.1177/0265532209346454
Xi, X. (2010b). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*, 291–300. https://doi.org/10.1177/0265532210364643
Xi, X. (2010c). How do we go about investigating test fairness? *Language Testing, 27*, 147–170. https://doi.org/10.1177/0265532209349465
Xi, X., Bridgeman, B., & Wendler, C. (2013). Tests of English for academic purposes in university admissions. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 318–337). Malden, MA: Wiley-Blackwell.
Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater*SM *v1.0* (Research Report No. RR-08-62). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02148.x
Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing, 29*, 371–394. https://doi.org/10.1177/0265532211425673
Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL academic speaking test* (Research Report No. RR-06-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2006.tb02013.x
Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT™ speaking section and what kind of training helps?* (Research Report No. RR-09-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x
Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*, 1222–1255. https://doi.org/10.1111/j.1467-9922.2011.00667.x
Xi, X., Schmidgall, J., & Wang, Y. (2016). Chinese users' perceptions of the use of automated scoring for a speaking practice test. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English: Language constructs, consequences and conundrums* (pp. 150–175). New York, NY: Palgrave McMillan. https://doi.org/10.1057/9781137449788_8
Xu, X., Ke, F., & Lee, S. (2016). EVALUATING TEACHING COMPETENCY IN A 3D E-LEARNING ENVIRONMENT USING A SMALL-SCALE BAYESIAN NETWORK. *Quarterly Review of Distance Education, 17*(3), 61–74.
Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (Research Report No. RR-95-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01637.x
Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

Yang, Y. (2017). Test anxiety analysis of Chinese college students in computer-based spoken English test. *Journal of Educational Technology & Society, 20*(2), 63–73.
Yashima, T., Nishida, R., & Mizumoto, A. (2017). Influence of learner beliefs and gender on the motivating power of L2 selves. *The Modern Language Journal, 101*(4), 691–711.
Ye, L. (2013). A needs-based analysis of cross-cultural competence: A case study on spoken English learning experience of Chinese International Teaching Assistants in the US. *Journal of English as an International Language, 8*(2), 30–49.
Yook, E. L. (1999). An Investigation of Audience Receptiveness to Non-native Teaching Assistants. *Journal of the Association for Communication Administration (JACA), 28*(2), 71–77.
Yook, E. L., & Albert, R. D. (1999). Perceptions of international teaching assistants: the interrelatedness of intercultural training, cognition, and emotion. *Communication Education, 48*, 1–17. https://doi.org/10.1080/03634529909379148
Yoon, S.-Y., Pierce, L., Huensch, A., Juul, E., Perkins, S., Sproat, R., & Hasegawa-Johnson, M. (2009). Construction of a rated speech corpus of L2 learners' spontaneous speech. *CALICO Journal, 26*, 662–673. https://doi.org/10.1558/cj.v26i3.662-673
Young, J. W., Morgan, R., Rybinski, P., Steinberg, J., & Wang, Y. (2013). *Assessing the test information function and differential item functioning for the TOEFL Junior® Standard test* (Research Report No. RR-13-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02324.x
Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., . . ., Fang, L. (2017). *Preparing for the speaking tasks of the TOEFL iBT test: An investigation of the journeys of Chinese test takers* (Research Report No. RR-17-19). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12145
Yule, G., & Hoffman, P. (1990). Predicting success for international teaching assistants in a U.S. university. *TESOL Quarterly, 24*, 227–243. https://doi.org/10.2307/3586900
Yule, G., & Hoffman, P. (1993). Enlisting the help of U.S. undergraduates in evaluating international teaching assistants. *TESOL Quarterly, 27*, 323–327. https://doi.org/10.2307/3587154
Zahedi, K., & Shamsaee, S. (2012). Viability of construct validity of the speaking modules of international language examinations (IELTS vs. TOEFL iBT): Evidence from Iranian test-takers. *Educational Assessment, Evaluation & Accountability, 24*, 263–277. https://doi.org/10.1007/s11092-011-9137-z
Zahedkazemi, E. (2015). Construct validation of TOEFL iBT (as a conventional test) and IELTS (as a task-based test) among Iranian EFL test-takers' performance on speaking modules. *Theory and Practice in Language Studies, 5*, 1513–1519. https://doi.org/10.17507/tpls.0507.27
Zareva, A. (2005). What is new in the new TOEFL iBT 2006 test format? *Electronic Journal of Foreign Language Teaching, 2*(2), 45–57.
Zechner, K., Bejar, I. I., & Hemat, R. (2007). *Toward an understanding of the role of speech recognition in nonnative speech assessment* (Research Report No. RR-07-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02044.x

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*, 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Zechner, K., Yoon, S.-Y., Bhat, S., & Leong, C. W. (2017). Comparative evaluation of automated scoring of syntactic competence of non-native speakers. *Computers in Human Behavior, 76*, 672–682. https://doi.org/10.1016/j.chb.2017.01.060

Zhang, Y. (2008). *Repeater analyses for TOEFL iBT* (Research Memorandum No. RM-08-05). Princeton, NJ: Educational Testing Service.

Zhang Hill, Y., & Liu, O. L. (2012). *Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT reading performance?* (Research Report No. RR-12-22). Princeton, NJ: Educational Testing Service.

Zhao, C. G. (2013). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing, 30*, 201–230. https://doi.org/10.1177/0265532212456965

Zhao, C. G. (2017). Voice in timed L2 argumentative essay writing. *Assessing Writing, 31*, 73–83. https://doi.org/10.1016/j.asw.2016.08.004

Zheng, X. (2017). Translingual Identity as Pedagogy: International Teaching Assistants of English in College Composition Classrooms. *Modern Language Journal, 101*, 29–44. https://doi.org/10.1111/modl.12373

Zhou, J. (2009). What Is Missing in the International Teaching Assistants Training Curriculum? *Journal of Faculty Development, 23*(2), 19–24.

Zu, J., Moulder, B., & Morgan, R. (2017). A field test study for the TOEFL Primary reading and listening tests. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 99–117). New York, NY: Routledge.

Zwick, R., & Thayer, D. T. (1995). *A comparison of the performance of graduate and undergraduate school applicants on the Test of Written English* (Research Report No. RR-95-15). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01650.x